UNITED STATES PATENT APPLICATION

FOR

MANAGING NETWORK CONGESTION USING DYNAMICALLY ADVERTISED CONGESTION STATUS

INVENTORS:

Dean Cheng
Snehal Karia

MANAGING NETWORK CONGESTION USING DYNAMICALLY ADVERTISED CONGESTION STATUS

Field of the Invention

This invention relates to computer networks. In particular, the invention

5 relates to congestion management.

The Background of the Invention

Traffic congestion in a network causes many problems including disruption of services, delayed transmission, and system outage. The congestion

10 may occur when a particular node in the network becomes saturated or overloaded with connections and service calls.

A network node may be come congested for a number of reasons. First, a flood of packets may transit through the node during some peak hours due to high demand, causing congestion. Second, the node may have resource

15 problems such as memory capacity, bandwidth unavailability. Third, the node may undergo some maintenance mode switchover where resumption of calls would take longer that the stipulated outage time.

One technique to address the congestion problem is to drop or reject new call requests. This technique imposes a burden on the already congested node to

20 execute a task to process the calls before rejecting them. Consequently, there are numerous crankbacks in the network, resulting in degraded performance and loss of revenue for the carrier. For service-oriented applications such as high availability applications, this congestion is undesirable.

1

SUMMARY OF THE INVENTION

A method and apparatus are described for managing congestion in a network. For a receiving node, a congestion status associated with a node in the network is determined. The congestion status is advertised to at least one other node in the network. For a sending node, a congestion status associated with a receiving node in the network is received. The congestion status corresponds to a measured node condition at the receiving node. A call is routed to the receiving node based on the received congestion status.

Other features and advantages of the invention will be apparent from the detailed description and drawings provided herein.

2

BRIEF DESCRIPTION OF THE DRAWINGS

The present invention is illustrated by way of example and not limitation in the figures of the accompanying drawings in which like references indicated similar elements which:

5   Figure 1 shows a single peer system in which one embodiment of the invention can be practiced.

Figure 2 shows a hierarchical system in which one embodiment of the invention can be practiced.

Figure 3 shows a computer system for the congestion management.

10   Figure 4 shows a flowchart for a process to advertise the congestion status.

Figure 5 shows a flowchart for a process to respond to the advertised congestion status.

DETAILED DESCRIPTION

A method and apparatus are described for managing congestion in a network. For a receiving node, a congestion status associated with a node in the network is determined. The congestion status is advertised to at least one other node in the network. For a sending node, a congestion status associated with a receiving node in the network is received. The congestion status corresponds to a measured node condition at the receiving node. A call is routed to the receiving node based on the received congestion status.

The receiving node may be a transit node or a terminating node. The node may be a logical node which corresponds to a peer group of nodes in a hierarchical network. The call is routed to the node if the node is a terminating node or if the node is a transit node and the congestion status indicates that the node is not congested.

The advantages of the present invention include reducing network congestion, allowing the network to dynamically adapt to the changing conditions, reducing the outage of the call service, minimizing impact on users and support personnel, increasing revenue for the carrier, balancing traffic patterns, and eliminating traffic bottlenecks in the network.

In the following, the description refers to the Asynchronous Transfer Mode (ATM) model and the Peripheral Component Interconnect (PCI) bus as an interface example. It is contemplated that the technique is applicable to other models, buses, or network architectures with similar characteristics.

Figure 1 shows a single peer system 100 in which one embodiment of the invention can be practiced. The system 100 includes nodes N1 110, N2 120, N3 130, N4 140, N5 150, N6 160, N7 170, N8 180, customer premises equipment (CPE) 111, 112, 131, 132, 171, 172, 181, 182, and 183. The single peer system 100 represents a network in which nodes are interconnected at the same hierarchical

level and form a group. In one embodiment, the network is an ATM network having an interconnection model of the private network-to-network interface (PNNI).

Each of the nodes N1 110, N2 120, N3 130, N4 140, N5 150, N6 160, N7 170, and N8 180 is an ATM switch that performs switching and routing functions. A connection is made when a node requests a switched virtual circuit (SVC/ SPVC) call. Messages are sent and forwarded from one node to another via established connection links.   For example, node N1 110 is connected to nodes N2 120 and CPE's 111 and 113; node N6 160 is connected to nodes N3 130, N5 150, and N7 170. Each of the nodes N1 110, N2 120, N3 130, N4 140, N5 150, N6 160, N7 170, and N8 180 is capable of measuring its own operational conditions such as traffic flow status, resource availability, maintenance status, etc. The measurement can be performed by any method suitable for the nodes. This is typically done locally at each ATM switch or network node. For example, the measurement can be performed using inter-switch network information or Service Specific Connection Oriented Protocol (SSCOP) L3 as specified in the ATM UNI 3.1 and 3.0. The measured conditions are used to indicate a congestion status which indicates whether or not a node has become congested. This congestion status can be broadcast or advertised to other nodes within the network. The broadcasting or advertising of the congestion status can be performed by setting a transit flag in the node. This transit flag is accessible to other nodes. In one embodiment, the transit flag is one of a topology state parameter in a PNNI system. The topology state parameter is part of a PNNI topology state element (PTSE) which is transmitted in a PNNI topology state packet (PTSP). The PTSE is routing information that is flooded in a peer group. The PTSP contains one PTSE. The topology state parameters include metrics and attributes. Examples of the metrics are maximum cell transfer delay (MCTD),

5

maximum cell delay variation (MCDV), maximum cell loss ratio (MCLR), and administrative weight. Examples of attributes are available cell rate (ACR), cell rate margin (CRM), variation factor (VF), branching flag, and restricted transit flag.

5   Each of the nodes N1 110, N2 120, N3 130, N4 140, N5 150, N6 160, N7 170, and N8 180 may be connected to a CPE such as a workstation, a computer system, or a peripheral device. As illustrated in Figure 1, node N1 is connected to CPE 111 and 112, node N3 130 is connected to CPE 131 and 132, node N7 is connected to CPE 171 and 172, and node N8 180 is connected to CPE 181, 182,

10   and 183.

A node may be a transit node or a terminating node. A transit node is one through which a message is routed but is not a final destination. A terminating node is a destination node and is connected to at least one CPE. Each of the nodes has a congestion manager 105 to manage congestion at the node. The

15   topology shown in Figure 1 is for illustrative purposes only. Other network topologies and/or configurations are possible. Each of the ATM switches or nodes N1 110, N2 120, N3 130, N4 140, N5 150, N6 160, N7 170, and N8 180 is configured to have a congestion manager 105.

Figure 2 shows a hierarchical system 200 in which one embodiment of the

20   invention can be practiced. The hierarchical system 200 includes two hierarchical levels 201 and 202. The level 201 includes logical nodes A 210, B 220, and C 230. The level 202 includes nodes 211, 212, 213, 214, 221, 222, 223, 224, 225, 231, 232, and 233. The congestion management for the hierarchical system 200 is essentially similar to that of the peer group shown in Figure 1

25   except that the management is performed at the logical level.

Each of the logical nodes A 210, B 220, and C 230 corresponds to a peer group at the next lower level, i.e., level 202. The logical node A 210 corresponds

6

to a peer group including nodes 211, 212, 213, and 214. The logical node B 220 corresponds to a peer group including nodes 221, 222, 223, 224, and 225. The logical node C 230 corresponds to a peer group including nodes 231, 232, and 233. A logical node acts on the behalf of its child peer group. Each of the logical

5   nodes A 210, B 220, and C 230 has a congestion manager 105 to manage congestion at the corresponding peer group.

In a hierarchical system, SVC/ SPVC connections can cross multiple peer groups. For example, a SVC/ SPVC connection may go from the logical node A 210 to the logical node C 230 passing through the logical node B 220. Each of the

10   logical nodes represents its corresponding child peer group and manages the congestion of the peer group. For example, if the traffic condition at the peer group B 220 which includes nodes 221, 222, 223, 224, and 225, becomes congested, the parent logical node B220 advertises the congestion status to other logical nodes by setting its transit flag. The transit flag of each logical node is

15   accessible to other logical nodes. In one embodiment, the transit flag is one of a topology state parameter in a PNNI system. The topology state parameter is part of a PNNI topology state element (PTSE) which is transmitted in a PNNI topology state packet (PTSP).

Other peer groups receive the congestion status of the logical node B 220

20   and avoid routing SVC/ SPVC calls traversing the peer group of the logical node B 220. The congestion at the peer group B 220 is therefore reduced.

Figure 3 shows a computer system 300 for the congestion management. The computer system 300 may be used as part of an ATM switch, a host machine, a workstation, a local area network (LAN), and any other system or

25   subsystem connected to the network. The computer system 300 includes a processor 305, a host bus 310, a host bridge chipset 320, a system memory 330, a

peripheral bus 340, a mass storage device 350, a network interface device 355, and K peripheral devices 360_1 to 360_K.

The processor 305 represents a central processing unit of any type of architecture, such as complex instruction set computers (CISC), reduced

5   instruction set computers (RISC), very long instruction word (VLIW), explicitly parallel instruction set computing (EPIC), or hybrid architecture. The invention could be implemented in a multi-processor or single processor computer system.

The host bridge chipset 320 includes a number of interface circuits to allow the host processor 305 access to the system memory 330 and the peripheral

10   bus 340. The host bridge chipset 320 may include a memory controller, a bus interface circuit, and an I/O controller. The memory controller provides an interface to the system memory 330. The I/O controller provides control of I/O functions.

The system memory 330 represents one or more mechanisms for storing

15   information. For example, the system memory 330 may include non-volatile or volatile memories. Examples of these memories include flash memory, read only memory (ROM), or random access memory (RAM). The system memory 330 contains a program 332, a data storage 334, and the congestion manager 105 as shown in Figure 1 and Figure 2. Of course, the system memory 330 preferably

20   contains additional software (not shown), which is not necessary to understanding the invention.

The peripheral bus 360 provides bus interface to the mass storage device 350, the network interface 355, and the peripheral devices 360_1 to 360_K. In one embodiment, the peripheral bus 360 is the peripheral component interconnect

25   (PCI) bus.

The mass storage device 350 include CD ROM, floppy diskettes, and hard drives. The mass storage device 350 stores non-volatile information such as

programs or data. The mass storage device 350 provides a mechanism to read machine-readable media. When implemented in software, the elements of the present invention are essentially the code segments to perform the necessary tasks. The program or code segments can be stored in a processor readable

5   medium or transmitted by a computer data signal embodied in a carrier wave, or a signal modulated by a carrier, over a transmission medium. The "processor readable medium" may include any medium that can store or transfer information. Examples of the processor readable medium include an electronic circuit, a semiconductor memory device, a ROM, a flash memory, an erasable

10   ROM (EROM), a floppy diskette, a compact disk CD-ROM, an optical disk, a hard disk, a fiber optic medium, a radio frequency (RF) link, etc. The computer data signal may include any signal that can propagate over a transmission medium such as electronic network channels, optical fibers, air, electromagnetic, RF links, etc. The code segments may be downloaded via computer networks

15   such as the Internet, Intranet, etc.

The network interface device 355 provides an interface to a network such as ATM, LAN, WAN, etc. The peripheral devices 360_1 to 360_K may include an audio device, a multimedia device, a modem, a printer controller, etc.

Figure 4 shows a flowchart for a process 400 to advertise the congestion

20   status. The process 400 is used for a receiving node. The node may be a physical node or a logical node. A logical node acts on the behalf of its child peer group as described in Figure 2.

Upon START, the process 400 determines a congestion status at the node (Block 410). This determination can be performed by measuring a node

25   condition. The node condition may include a traffic condition, resource availability such as memory or processor, and maintenance status. Then, the process 400 determines if the congestion status indicates a congestion at the node

(Block 420). If there is not congestion, the process 400 resets a "transit restricted"

flag indicating that the node is not restricted for transit (Block 440). This transit

flag is accessible to other nodes in the network. If there is a congestion, the

process 400 sets a "transit-restricted" flag to indicate that all calls through the

5   node should be avoided unless the node is a terminating node (Block 430).

Next, the process 400 advertises the congestion status by making the flag

available for access to at least one other physical or logical node in the network

(Block 450). Then, the process 400 is terminated.

Figure 5 shows a flowchart for a process 500 to respond to the advertised

10   congestion status. The process 500 is used for a sending node. The node may be a

physical node or a logical node. A logical node acts on the behalf of its child peer

group as described in Figure 2.

Upon START, the process 500 receives a congestion status associated with

a receiving node (Block 510). This congestion status corresponds to a measured

15   node condition at the receiving node. Typically, the receiving of the congestion

status is performed by accessing a transit flag of the receiving node. Then, the

process 500 determines if the node is a termination node (Block 520). If the

receiving node is a terminating node, the process 500 routes the SVC/ SPVC call

to the node (Block 550). The process 500 is then terminated.

20   If the receiving node is not a terminating node, the process 500 determines

if the congestion status indicates that there is a congestion at the node (Block 530).

If there is no congestion, the process 500 goes to block 550. If there is a

congestion, the process 500 routes the SVC/ SPVC call to another receiving node.

Then the process 500 is terminated.

25   A technique has been described to manage congestion in a network. For a

receiving node, a congestion status associated with a node in the network is

determined. The congestion status is advertised to at least one other node in the

network. For a sending node, a congestion status associated with a receiving node in the network is received. The congestion status corresponds to a measured node condition at the receiving node. A call is routed to the receiving node based on the received congestion status.

5   In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereto without departing from the broader spirit and scope of the invention as set forth in the appended claims. The specification and drawings are, accordingly, to be

10   regarded in an illustrative rather than a restrictive sense.